2018. *Proc Ling Soc Amer* 3. 34:1-15. https://doi.org/10.3765/plsa.v3i1.4322

Experimental support for a one-step model of phoneme acquisition

Emily Moeng*

Abstract. Phonetic category acquisition involves a distributional learning mechanism (Maye, Werker, and Gerken 2002). Some researchers suggest that phonetic category acquisition is only the first step in a two-step model of phonological acquisition by modelling these two steps separately (Guenther and Gjaja, 1996; Boersma, Escudero, and Hayes, 2003; Peperkamp, Pettinato, and Dupoux 2003; Peperkamp, Calvez, Nadal, and Dupoux 2006), while others have argued for a one-step model (Dillon, Dunbar, and Idsardi 2013). This experimental study maps the learning trajectory of three groups of adult learners: (1) a group exposed to a bimodal frequency distribution where both halves of the bimodal distribution occur in complementary environments (Bimodal-Comp group), (2) a group exposed to a bimodal frequency distribution where both halves of the bimodal distribution occur in non-complementary environments (Bimodal-NonComp group), and (3) a group exposed to a monomodal frequency distribution (Monomodal group). This study finds support for a one-step model of phoneme acquisition, with the Bimodal-Comp group having lower sensitivities to critical stimuli than even the Monomodal group at all three exposure times tested.

Keywords. distributional learning; acquisition; artificial language; phoneme acquisition; phonetic category; phonetic category acquisition; Mechanical Turk

1. Introduction. At least two theories regarding the acquisition of sound categories have been proposed: a one-step model, and a two-step model. In a two-step model, learners initially acquire phonetic categories by noting clusters of high-frequency[1] distributions (Maye and Gerken 2002). If these phonetic categories occur in predictably distinct environments from other similar-sounding phonetic categories, learners in a second step will learn that these are allophones of a single phoneme (Peperkamp et al. 2006). However in a one-step model, learners search for subsets of sets, where subsets within a set are in predictably distinct environments from other subsets within the same set (Dillon, Dunbar, and Idsardi 2013). In this way, learners acquire allophones and the rules relating allophones to one another in a single step.

This study maps the learning trajectory of three groups of learners: one group exposed to a bimodal frequency distribution where both halves of the bimodal distribution occur in complementary environments (the **Bimodal-Comp** group), one group exposed to a bimodal frequency distribution where both halves of the bimodal distribution occur in non-complementary environments (the **Bimodal-NonComp** group), and one group exposed to a monomodal frequency distribution (the **Monomodal** group). Participants were placed in one of three ExposureTime groups: One, Two, or Three. Participants in ExposureTime One heard one block of training items during the Training phase (which consisted of 96 critical tokens and lasted approximately 5 minutes); participants in ExposureTime Two heard two blocks of training items during Training (192 critical tokens, 10 minutes); and participants in ExposureTime Three heard three blocks of

* Many thanks to the P-side Research Group at the University of North Carolina at Chapel Hill for feedback on this study. All errors are my own. Author: Emily Moeng, University of North Carolina at Chapel Hill (e.moeng@gmail.com).

[1] By "frequency," we refer to count frequency, not to auditory frequency as measured in Hertz.

training items (288 critical tokens, 15 minutes). This study finds support for a one-step model of phoneme acquisition, with the Bimodal-Comp group having: 1) numerically lower sensitivities to critical stimuli than both the Monomodal and Bimodal-NonComp groups at all three times tested, 2) significantly lower sensitivities than the Monomodal group at ExposureTime Two, and 3) significantly lower sensitivities to critical stimuli than the Bimodal-NonComp group at ExposureTime Three.

2. Background. When determining a language's phonemic inventory, a linguist may approach the problem by first noting phones in the language, and subsequently determining if any of these phones are variant pronunciations of a single phoneme. Although phones are typically treated as being wholly acoustic, phones themselves are made up of a group of sounds with variant pronunciations, therefore requiring "phones" to refer to categories that must also be acquired by language learners. This sound category has been referred to as a "phonetic category" (for example, Werker, Pons, Dietrich, Kajikawa, Fais, and Amano 2007; Dillon et al., 2013, Maye and Gerken, 2000) or "phonetic equivalence class" (Maye and Gerken, 2000) and varies by language. For example, both English-learning and Mandarin-learning infants will hear tokens of [i] and [y]. The English-learning infant may hear [i] in a word like ***ea****ts*; the Mandarin-learning infant may hear [i] in a word like 筆 [bi] 'pen'; the English-learning infant may hear [y] following a rounded context as in the phrase *Lou* ***ea****ts*; the Mandarin-learning infant may hear [y] in a word like 女 [ny] 'woman.' However despite being exposed to both phones, the English-learning infant must learn that [i] and [y] are variant pronunciations of a single phonetic category, whereas the Mandarin-learning infant must learn that they belong to two different phonetic categories (see Werker, Yeung, and Yoshida 2012). The acquisition of phonetic categories has been noted to occur in infants anywhere between the age of 6 months (Kuhl, Williams, Lacerda, Stevens, and Lindblom 1992) to 10 months of age (Werker and Tees 1984; Werker, Gilbert, Humphrey, and Tees 1981; Eilers, Gavin, and Wilson 1979; Eimas, Siqueland, Juscyk, and Vigorito 1971), and has also been found in adult participants of artificial language learning studies (Maye and Gerken 2000; Feldman, Griffiths, and Morgan 2009; Maye and Gerken 2001; Escudero, Benders, and Wanrooij 2011). Researchers such as Maye et al. (2002) and Werker et al. (2012) argue that language learners acquire phonetic categories through **distributional learning**. A distributional learning account of sound acquisition claims that language learners note statistical distributions of tokens that they have been exposed to and make inferences about the phonetic categories in the language they are being exposed to from these distributions (Maye et al. 2002). Learners exposed to a bimodal distribution of tokens along some phonetic dimension(s) will infer that there are two phonetic categories, whereas learners exposed to a monomodal distribution will infer that there is only one phonetic category.

Distributional learning is widely cited as a mechanism for phonetic category acquisition (for example, Kuhl 2004; Kuhl, Stevens, Hayashi, Deguchi, Kiritani, Iverson 2006; Werker et al. 2012), and has been experimentally supported in artificial language learning tasks for both adults (Maye and Gerken 2000; Maye and Gerken 2001; Hayes-Harb 2007; Escudero et al. 2011) and infants (Maye et al. 2002). Maye and Gerken (2000) find that participants who are exposed to a bimodal distribution of critical tokens are more likely to respond that tokens taken from opposite ends of the bimodal distribution are "different" compared to participants exposed to a monomodal distribution of critical tokens. In a similar study with a different analysis, Hayes-Harb (2007) finds that participants exposed to a bimodal distribution of critical tokens have higher sensitivities, measured in *d'*, than participants exposed to a monomodal distribution of critical tokens.

Attempts to replicate Maye and Gerken's (2000) findings with other stimuli have shown mixed success. Stimuli successfully used in replications include the stop pairs [t] vs. [d], and [k] vs. [g] (Maye and Gerken 2001, Maye et al. 2002, Hayes-Harb 2007); the vowel pairs [a] vs. [ɑ], and [i] vs. [ɪ] (Gulian, Escudero, and Boersma 2007; Escudero et al. 2011); and the Thai tone pairs [33] and [241] (Ong, Burnham, and Escudero 2015). However, Peperkamp et al. (2003) failed to replicate these findings when testing fricatives ranging from [ʁ] to [χ] with French-speaking adult participants.

2.1. LEARNING COMPLEMENTARY DISTRIBUTION. While there have been numerous replications of Maye and Gerken's distributional learning experiment, there have been far fewer experimental studies which have studied whether learners can learn that two allophones belong to a single phoneme. Two artificial language learning experiments which test whether participants can learn that two sounds are in complementary distribution are Peperkamp et al. (2003) and a recent dissertation, Noguchi (2016).

Peperkamp et al. (2003) tested three groups of French speakers: a Monomodal group, a Bimodal group, and a Bimodal+Assimilation group. VC target syllables were created, where V consisted of one of the three vowels [i a u], and C consisted of an 8-point continuum between the fricatives [ʁ] and [χ]. These were followed by CV context syllables, which began with either a voiced or voiceless consonant, creating $VC_{Target}.CV_{Context}$ "phrases." The Monomodal group heard a monomodal distribution of the fricatives [ʁ] and [χ] during the exposure phase, whereas both Bimodal groups heard them in a bimodal distribution. The Bimodal+Assimilation group only heard the [ʁ]-half of the continuum before voiced consonants, and the [χ]-half of the continuum before voiceless consonants. During the test phase, participants were presented with pairs of 2-word VC.CV "phrases," and were asked whether the first words in these two phrases were the same or different. This test phase occurred once before the exposure phase, and once after. Peperkamp and colleagues found that the Bimodal group was the only group to show a significant difference between the pre- and post-test phases, but found no significant interaction across groups. Note that for this study, the participants as native speakers of French already have the phonological rule specified in the Bimodal+Assimilation group, so the Monomodal and Bimodal groups also had experience with this assimilation rule through their L1.

In a recent dissertation, Noguchi (2016) tested three groups of native English speaking adults: a Non-Complementary group, a Complementary group, and a Control group. The first two groups heard a bimodal distribution of critical syllables with the onset ranging from an alveopalatal fricative [ɕɑ] to a retroflex fricative [ʂɑ] in an 8-point continuum. (The Control group did not hear any of the critical syllables containing fricatives.) The Non-Complementary group heard all 8 points of the continuum following one of four context syllables, all of which ended with [i], and also all 8 points of the continuum following one of four context syllables, all of which ended with [u] (e.g. [liɕɑ], [liʂɑ], [luɕɑ], and [luʂɑ]). The Complementary group only heard the four tokens on the [ɕɑ]-side of the 8-point continuum (referred to here as S_1a-S_4a) following syllables ending with [i], and the four tokens on the [ʂɑ]-side of the 8-point continuum (referred to here as S_5a-S_8a) following syllables ending with [u] (e.g. [liɕɑ] and [luʂɑ]). Subsequently, participants were tested on whether they believed the syllables presented in isolation were the "same" or "different" from one another. Noguchi found that the Complementary group had lower sensitivity (lower *d'*) than the Control and the Complementary groups. Noguchi interprets this result as showing that the Complementary group treated [ɕ] and [ʂ] as allophones of the same phoneme.

2.2. A ONE- OR TWO-STEP MODEL OF PHONEME ACQUISITION. The tradition of studying the acquisition of phonemes which consist of multiple phonetic categories in isolation from the study of phonetic category acquisition has carried on under the assumption that phonetic categories are formed before language learners form phonemes which consist of multiple phonetic categories (see Dillon et al., 2013). For example, Peperkamp et al. (2006) seem to suggest that language learners construct phonemes in two steps by assuming an input of phonetic categories when modelling the acquisition of allophonic rules. Therefore, it is assumed that phonetic categories have been established before phonemes are acquired.

However, in a modelling study of Inuktitut vowels, Dillon et al. (2012) argue that a two-step model of phoneme acquisition is not a probable method utilized by language learners given the large amount of category overlap exhibited across phonemes. Inuktitut contains three vowel phonemes: /i/, /u/, and /a/. Each of these phonemes consists of two allophones: respectively [i u a] which occur after non-uvulars, and their lowered vowel counterparts [e o ɑ] which occur after uvulars. In a two-step model, the learner must discover six phonetic categories in an initial step, then determine that, for example, [i] and [e] occur in complementary environments and therefore are allophones of a single phoneme.

In a clustering analysis of Inuktitut vowels, Dillon et al. show that a machine learner performs poorly if tasked with determining the six allophones of Inuktitut. They show that a simple mixture of Gaussians model either discovers too few allophones, or discovers clusters which are not accurate enough to actual phonetic categories for learners to then determine that these categories are in complementary environments with other categories in a second step. Because of this, Dillon et al. suggest a one-step model of phonological acquisition, in which allophones and rules relating allophones to one another are acquired in a single step. In their model, learners search for subsets of sets, under the condition that subsets have the same parameters as one another, and are in complementary distribution with other subsets within their set. They find that modelling with a multivariate mixture of linear models is more accurate in approximating the allophones and phonemes of Inuktitut compared to a simple mixture of Gaussians model which discovers allophones, and which would then need to be followed by some secondary step to discover phonemes.

The artificial language learning experiment described here will test for whether we find evidence for a two-step model or a one-step model by mapping the learning trajectory of three groups of learners. This study finds support for a one-step model.

3. Research question. This study will have two goals. The first is to replicate the results of Noguchi (2016) through Mechanical Turk. Noguchi found that, even after one day of training, a group trained on a bimodal distribution where S_1a-S_4a occurred after [i] and S_5a-S_8a occurred after [u] (Bimodal-Comp group) had significantly lower sensitivity (measured in *d'*) than a group trained on a bimodal distribution where all tokens along the S_1a-S_8a continuum occurred after [i] and after [u] (Bimodal-NonComp group). This study will also include a monomodal group for comparison.

The second goal of this study is to determine whether there is experimental evidence for a one-step model of phoneme acquisition or a two-step model. In order to do so, this study will randomly place participants into one of three exposure times, with the greatest exposure period consisting of roughly the same number of critical tokens as that found in Noguchi's first day of training. Noguchi found a significant difference in *d'* between the Bimodal-Comp and Bimodal-NonComp groups after one day of training, so this study will use that point as roughly the last exposure time in order to determine how each group behaves before that point.

Based on previous studies comparing the effects of exposure to a bimodal (non-complementary) distribution and a monomodal (non-complementary) distribution (e.g. Hayes-Harb 2007), we expect the Bimodal-NonComp group to achieve a higher sensitivity than the Monomodal group over time (see orange and red lines in Figure 1). Based on Noguchi (2016), we expect the Bimodal-NonComp group to have a significantly higher sensitivity compared to the Bimodal-Comp group after exposure to at least 256 critical syllables during training (see orange and green points in Figure 1).

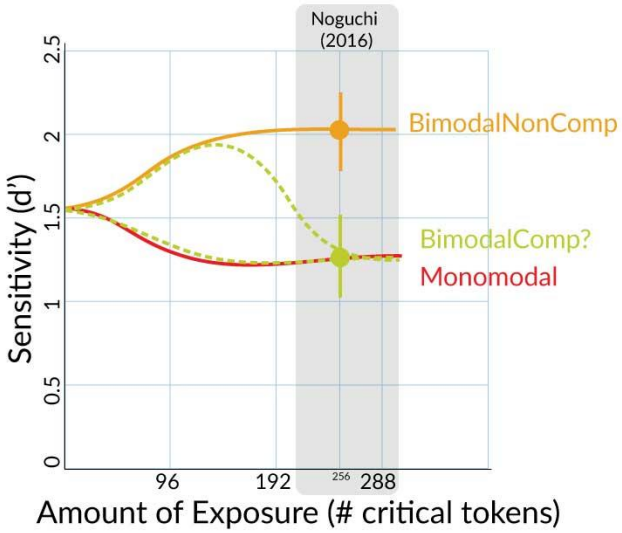

Figure 1. Predicted results as amount of exposure increases. BimodalNonComp and Bimodal-Comp points indicated at an exposure amount of 256 illustrate findings from Noguchi (2016).

The learning trajectory of interest is that of the Bimodal-Comp group. If phoneme acquisition follows a two-step model in which learners first acquire phonetic categories through distributional learning, and then learn that two phonetic categories are allophones of a single phoneme, we would expect the Bimodal-Comp group to initially pattern with the Bimodal-NonComp group, and only later pattern with the Monomodal group. If phoneme acquisition follows a one-step model of acquisition, in which learners are searching for subsets of sets from the very beginning, where each subset must be in complementary distribution with any other subsets in the same set, then the Bimodal-Comp group should always pattern with the Monomodal group. To summarize, the two research questions we ask in this paper are as follows:

(1) Can the results of Noguchi (2016) be replicated through an online platform such as Mechanical Turk?
(2) Do we find experimental evidence for a one- or a two-step model of phoneme acquisition?

This study successfully replicates Noguchi (2016) on Mechanical Turk, and finds numerical support for a one-step model of phoneme acquisition.

4. Methodology. This experiment closely followed the methodology of Noguchi (2016). The main differences between Noguchi and the current experiment are that (1) training phases lasted for one of three times (approximately 5, 10, or 15 minutes) in order to map the learning trajectory of each distribution type, (2) a rule test phase was included to determine whether there was

evidence that learners exposed to a complementary distribution of critical phones learned a rule, and (3) a group trained on a monomodal distribution was included.

4.1. STIMULI. Stimuli consisted of four types of syllables: critical syllables, filler syllables, context syllables, and generalization syllables. Following Noguchi (2016), onsets of critical syllables were drawn from an 8-point continuum ranging between an alveopalatal fricative [ɕ] to a retroflex fricative [ʂ]. Continuum points will be referred to as S_1-S_8, where S_1 indicates the most [ɕ]-like end of the continuum, and S_8 indicates the most [ʂ]-like end. Each onset was followed by [ɑ]. Filler syllables consisted of the syllables [tɑ] and [t^hɑ]. Context syllables were [pi pu hi hu ni nu], where each of the three onsets [p h n] are paired with either [i] or [u][2]. Generalization syllables also ended in either [i] or [u], but had different onsets [ti tu fi fu li lu k^hi k^hu mi mu ɹi ɹu].

All recordings, filtering, and splicing were done in Praat (version 6.0.29, Boersma, 2002), software for speech analysis, synthesis, and manipulation. Stimuli were recorded by the experimenter, a native speaker of English and heritage speaker of Mandarin. Recordings were made in a soundproof booth on an HP Spectre laptop at 44100 Hz using an ATR2500-USB Audio Technica microphone. The experimenter recorded all tokens embedded in two-syllable "phrases," with context syllables and generalization syllables occurring at the beginning of the phrase, and test syllables and filler syllables occurring at the end of the phrase. Context and generalization syllables were followed by a dummy syllable [ʃɑ] (e.g. [hi ʃɑ]); test syllables and filler syllables were preceded by a dummy syllable [ɑ] (e.g. [ɑ ɕɑ]). Before manipulations were made, all recordings were high-pass filtered for frequencies equal to or below 200 Hz. Dummy syllables were then spliced out. All cuts were made where the waveform crossed 0 Hz to avoid clicks and other unnatural non-speech sounds when splicing sounds together.

For critical syllables, tokens of the two endpoints of the target continuum [ɕɑ] and [ʂɑ] were recorded (again, spliced from an original recording of dummy-critical "phrases"). The fricative portions ([ɕ] and [ʂ]) and vowel portions ([ɑ]) were isolated. The middle 160 ms of each fricative was extracted using a parabolic windowing function. The mean intensity of each fricative was adjusted to 60 dB. To create the fricative portion of the 8-point continuum, the endpoint fricatives were overlapped in varying amounts, with the second point of the continuum consisting of 6/7ths of the [ɕ] token and 1/7th of the [ʂ] token, the third point of the continuum consisting of 5/7ths of the [ɕ] token and 2/7th of the [ʂ] token, etc. The continuum between vowels was created by using TANDEM-STRAIGHT, software which creates natural-sounding continua between two sounds. The vowel spliced from [ɕɑ] and the vowel spliced from [ʂɑ] were used as input into TANDEM-STRAIGHT (details of the process TANDEM-STRAIGHT uses to create continua can be found in Kawahara (2008)). TANDEM-STRAIGHT allows the user to mark any number of landmarks on one spectrogram (for example, the beginning of the steady state of the vowel, the onset of voicing, etc.) that corresponds to a similar landmark on another spectrogram, so that durations between landmarks can be stretched or compressed in the generated continuum points. TANDEM-STRAIGHT returned 6 intermediate stimuli, for a total of 8 continuum points including the endpoints. All vowel continuum points were then scaled to have a mean intensity of 72 dB. Following this, each of the 8 fricative sounds were spliced onto their corresponding 8 vowel sounds, creating an 8-point continuum between [ɕɑ] and [ʂɑ]. Each of these syllables was scaled to have an average intensity of 74 dB. Critical syllables will be referred to as S_1a-S_8a, where S_1a refers to the most [ɕɑ]-like end, and S_8a refers to the most [ʂɑ]-like end. Examples of critical syllables S_1a, S_3a, S_6a, and S_8a can be found in Figure 2.

[2] [t] and [p] here both refer to voiceless unaspirated stops.

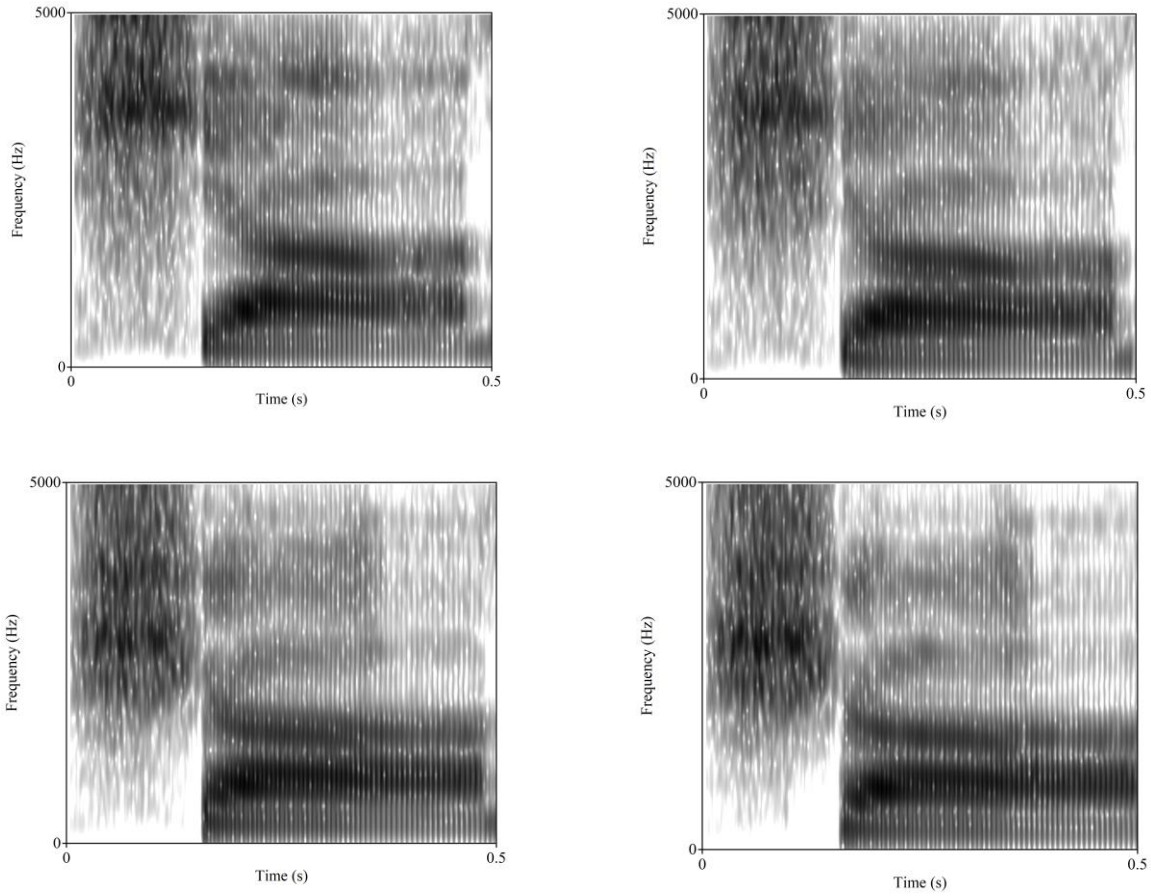

Figure 2. First 500 ms of critical syllables S_1a (top left), S_3a (top right), S_6a (bottom left), and S_8a (bottom right).

Each of the 6 context syllables was concatenated before each of the 8 critical syllables and each of the 4 filler syllables (for example, *hi* S_1a). This made up the stimuli to be used during training. Each of the 12 generalization syllables was concatenated before S_1a and S_8a, and also before each of the 4 filler syllables (for example, *li* S_1a). These stimuli were used in the rule test, which is described below.

4.2. PROCEDURE. Participants were randomly placed into one of three Distributions: Bimodal-Comp, Bimodal-NonComp, or Monomodal. Participants were also randomly placed into one of three ExposureTimes (One, Two, or Three), and one of two TestOrders (RuleFirst or PhoneFirst). This experiment consisted of five parts: a practice test in English, followed by a training phase, followed by two tests (a rule test and a phone test), and ending with a questionnaire.

During the **English phone practice test**, participants were given a pair of English words, such as *sheep* and *ship*. Participants were asked to determine whether these words were the same word, or two different words. Participants were given 4 same pairs (e.g. $ship_1$ vs. $ship_2$), and 4 different pairs (e.g. $ship_1$ vs. $sheep_1$). No feedback was given.

Following the practice phone test, participants were directed to a **training phase**. Before training, participants were told that they would hear two-word phrases in a langauge they had not heard before and were asked to listen passively to these phrases without writing anything down.

Phrases consisted of context syllables followed by critical syllables, or context syllables followed by filler syllables.

Participants in the Bimodal-Comp and Bimodal-NonComp groups were exposed to critical phones whose frequencies fell in a bimodal distribution, and participants in the Monomodal group were exposed to critical phones whose frequencies fell in a monomodal distribution. Figure 3 illustrates the frequency distributions participants were exposed to.

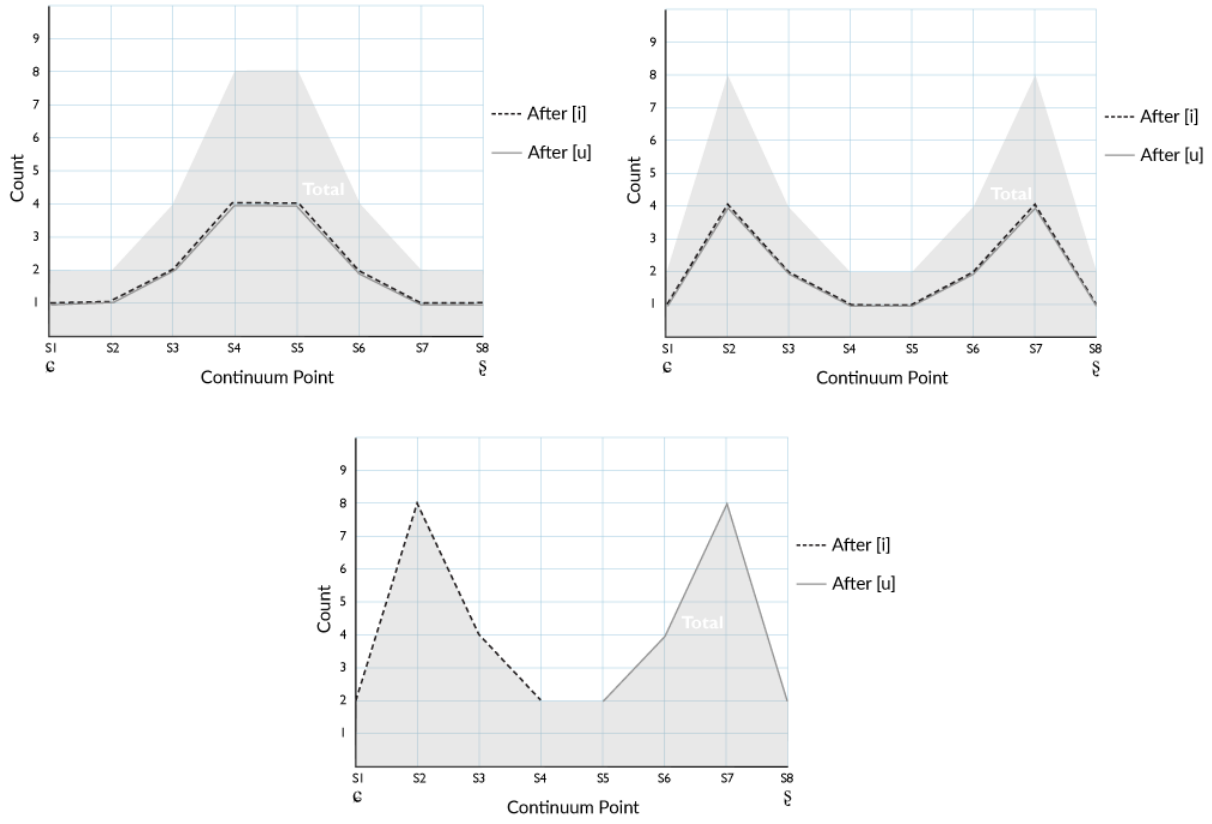

Figure 3. Illustration of familiarization frequency for Monomodal group (top left), Bimodal-NonComp group (top right), and Bimodal-Comp group (bottom), during training.

Each block of training consisted of one of each of the 6 context syllables followed by 16 critical syllables (following the distributions shown in Figure 3), resulting in 96 critical phrases per block. Each block of training also consisted of one of each of the 6 context syllables followed by one of the 2 filler syllables, resulting in 48 fillers per block. Participants in the ExposureTime One group were trained with one block of training stimuli (96 critical stimuli, 48 fillers); participants in the ExposureTime Two group were trained with two blocks of training stimuli (192 critica, 96 fillers); and participants in the ExposureTime Three group were trained with three blocks of training stimuli (288 critical, 144 fillers).[3] Training lasted about 5 minutes for participants in the ExposureTime One group, 10 minutes for the ExposureTime Two group, and 15 minutes for the ExposureTime Three group.

[3] For reference, participants in Noguchi (2016) heard 256 critical stimuli during training on Day 1 of his two-day study. Noguchi finds that his Bimodal-NonComp group had a significantly greater average *d'* than his Bimodal-Comp group at the end of training on Day 1.

After training, participants were directed to one of the two test phases. In the **phone test**, participants were presented with pairs of syllables. Participants were asked to determine whether the syllables in these pairs were the same or different. The phone test phase consisted of 12 different critical pairs (S_1a vs. S_8a), 12 same critical pairs (S_1a vs. S_1a or S_8a vs. S_8a), 12 different filler pairs ([ta]$_1$ vs. [t^{h}a]$_1$), and 12 same filler pairs ([ta]$_1$ vs. [ta]$_2$).

In the **rule test**, participants were told they would hear two phrases played, but only one of the phrases was allowed in the language they had just heard. Phrase pairs were either old phrase pairs, new phrase pairs, or filler phrase pairs. Old phrases were taken from those found in the training phase, and consisted of a context syllable followed by a critical or filler syllable (e.g. *hi* S_1*a*). New phrases consisted of a generalization syllable followed by a critical syllable (e.g. *li* S_1*a*). Filler phrases consisted of a generalization syllable followed by a filler syllable. For reasons of space, results of the rule test will not be discussed here.[4]

Participants in the PhoneFirst condition took the phone test first, followed by the rule test. Participants in the RuleFirst condition took the rule test first, followed by the phone test.

After the experiment, participants were directed to a short **questionnaire** which asked about participants' demographic information (age, place of residence, etc.), language background (native language, languages studied, history of speech or hearing disorder, etc.), and attention levels during participation. Participants were also asked whether they had noticed any patterns and whether they used any strategies during the experiment.

4.3. PARTICIPANTS. Participants were recruited on Mechanical Turk ("MTurk"), an online recruitment platform hosted by Amazon.com. Participants were asked to participate only if they (1) had no known history of speech or hearing impairments, (2) were a native speaker of English, (3) had regular access to a computer with an internet connection, and (4) were using a computer able to play audio. Because this experiment was conducted online rather than face-to-face, only participants using a computer in the United States were allowed to participate to increase the chance that the participant would be a native English speaker. This can be done through an MTurk "qualification," attributes that participants on MTurk can obtain. In addition, since the onsets of the critical syllables ranged between [ɕ] and [ʂ], following Noguchi (2016), participant responses were not included in analysis if they reported having some background in a language with more than one voiceless post-alveolar fricative as phonemes. Participants were also asked to not participate if they had some language background in Mandarin Chinese, Japanese, Russian, or German. Qualifications used to screen participants are as follows:

- Only Workers using a computer in the United States were allowed to participate
- Only Workers who had an approval rating of equal or greater to 90% on all tasks they had completed on MTurk ("HITs") were allowed to participate
- Only Workers who had at least 50 tasks approved by those putting forth tasks ("Requesters") were allowed to participate

489 participants were recruited through Mechanical Turk. Participants were excluded if they (1) scored fewer than 5/8 correct on the practice English test (15 excluded for this), (2) reported having a speech or hearing disorder in the questionnaire (4 excluded for this), (3) reported not being a native English speaker (1 excluded for this), (4) reported having some sort of background with a language with more than one voiceless post-alveolar fricative (22 excluded for this). This left

[4] For a full discussion, see Moeng (forthcoming).

the following number of participants per condition (note that some participants were excluded for more than one reason):

		First Timepoint	Second Timpoint	Third Timepoint
Bimodal-Comp	PhoneFirst	23	23	25
	RuleFirst	18	17	18
	Total	**41**	**40**	**43**
Bimodal-NonComp	PhoneFirst	31	17	19
	RuleFirst	20	16	21
	Total	**51**	**33**	**40**
Monomodal	PhoneFirst	28	23	26
	RuleFirst	20	21	25
	Total	**48**	**44**	**51**

Table 1: Number of participants per condition.

5. Results. Section 5.1 will describe the model used to analyze participant responses.

5.1. MODEL USED IN ANALYSIS. The regression used in analysis modelled one dependent variable, **Response**, with three fixed effects: (1) **Distribution**, consisting of three levels {*Bimodal-Comp, Bimodal-NonComp, Monomodal*}, (2) **PairType**, consisting of two levels {*SamePair*, *DiffPair*}, and (3) **ExposureTime**, consisting of three levels {*One*, *Two*, *Three*}. The dependent variable **Response** consists of two levels, *s* and *d*, where *s* corresponds to a participant response of "same" during the Test phase, and where *d* corresponds to a participant response of "different" during the Test phase. Random slopes by Subject and by Item are included. The regression was fitted to the formula in (3)[5]:

(3) Response ~ Distribution * PairType * ExposureTime + (1|Subject) + (1|Item)

Follow-up contrasts in the context of the overall model were completed to test the following hypotheses:

- The interaction between Distribution and PairType is significant for the Bimodal-Comp group compared to the Bimodal-NonComp group
- The interaction between Distribution and PairType is significant for the Bimodal-Comp group compared to the Monomodal group
- The interaction between Distribution and PairType is significant for the Bimodal-NonComp group compared to the Monomodal group

These three hypotheses were tested at each of the three ExposureTimes. Results are summarized in Table 2 (critical trials) and Table 3 (control trials).

[5] A model fitted to a formula with a more complex random effects structure (Response ~ Distribution*PairType*ExposureTime + (1+Pairtype|Subject) + (1+Timepoint+Condition|Item)) failed to converge.

ExposureTime	Distribution Comparison	Coefficient	SE	Wald Z	*p*
One	Bimodal-Comp vs. Bimodal-NonComp	-0.003	0.383	-0.008	0.993
	Monomodal vs. Bimodal-Comp	-0.497	0.443	-1.121	0.262
	Monomodal vs. Bimodal-NonComp	-0.494	0.405	-1.220	0.223
Two	Bimodal-Comp vs. Bimodal-NonComp	-0.146	0.372	-0.392	0.695
	Monomodal vs. Bimodal-Comp	-0.831	0.403	-2.060	0.039 *
	Monomodal vs. Bimodal-NonComp	-0.685	0.424	-1.614	0.106
Three	Bimodal-Comp vs. Bimodal-NonComp	-0.906	0.391	-2.316	0.021 *
	Monomodal vs. Bimodal-Comp	-0.494	0.343	-1.442	0.149
	Monomodal vs. Bimodal-NonComp	0.411	0.387	1.063	0.288

Table 2: Summary of follow-up contrasts testing specific hypotheses for critical trials.

At ExposureTime Two, the interaction between Condition and PairType when comparing the Bimodal-Comp group with the Monomodal group is significant for critical trials ($p = 0.039$). At ExposureTime Three, the interaction between Condition and PairType when comparing the Bimodal-Comp group with the Bimodal-NonComp group is significant for critical trials ($p = 0.021$). There are no significant interactions between Condition and PairType at ExposureTime One. We interpret these findings as follows: with the least amount of exposure tested in this experiment (ExposureTime One), no groups differ significantly from one another in terms of sensitivity. At the second-most amount of exposure tested in this experiment (ExposureTime Two), the Bimodal-Comp group and the Monomodal group do not pattern together in terms of sensitivity, with the Monomodal group having 0.697 greater difference in probability between responses in DiffPairs and SamePairs (where log-odds = 0.831 and probability = $e^{\text{log-odds}}/(1+e^{\text{log-odds}})$) compared to the difference in probability between responses in DiffPairs and SamePairs in the Bimodal-Comp group. At ExposureTime Three, the Bimodal-Comp and Bimodal-NonComp groups do not pattern together, with the Bimodal-NonComp group having 0.712 (0.906 log-odds) greater difference in probability between responses in DiffPairs and SamePairs compared to the difference in probability between responses in DiffPairs and SamePairs in the Bimodal-Comp group. This successfully replicates findings from Noguchi (2016), who finds that a Bimodal-Comp group has lowered sensitivity compared to a Bimodal-NonComp group.

ExposureTime	Distribution Comparison	Coefficient	SE	Wald Z	*p*
One	Bimodal-Comp vs. Bimodal-NonComp	0.125	0.325	0.385	0.700
	Monomodal vs. Bimodal-Comp	0.096	0.330	0.290	0.772
	Monomodal vs. Bimodal-NonComp	-0.030	0.308	-0.096	0.923
Two	Bimodal-Comp vs. Bimodal-NonComp	0.710	0.352	2.017	0.044 *
	Monomodal vs. Bimodal-Comp	-0.036	0.349	-0.103	0.918
	Monomodal vs. Bimodal-NonComp	-0.746	0.344	02.168	0.030 *
Three	Bimodal-Comp vs. Bimodal-NonComp	-0.035	0.365	-0.096	0.924
	Monomodal vs. Bimodal-Comp	-0.048	0.344	-0.140	0.889
	Monomodal vs. Bimodal-NonComp	-0.013	0.351	-0.038	0.970

Table 3: Summary of follow-up contrasts testing specific hypotheses for control trials.

For control trials, there was a significant interaction between Condition and PairType between the Bimodal-Comp and Bimodal-NonComp groups, as well as a significant interaction between the Monomodal and Bimodal-NonComp groups. It is unclear what resulted in the significantly lower sensitivity in control stimuli for the Bimodal-NonComp group at ExposureTime Two.

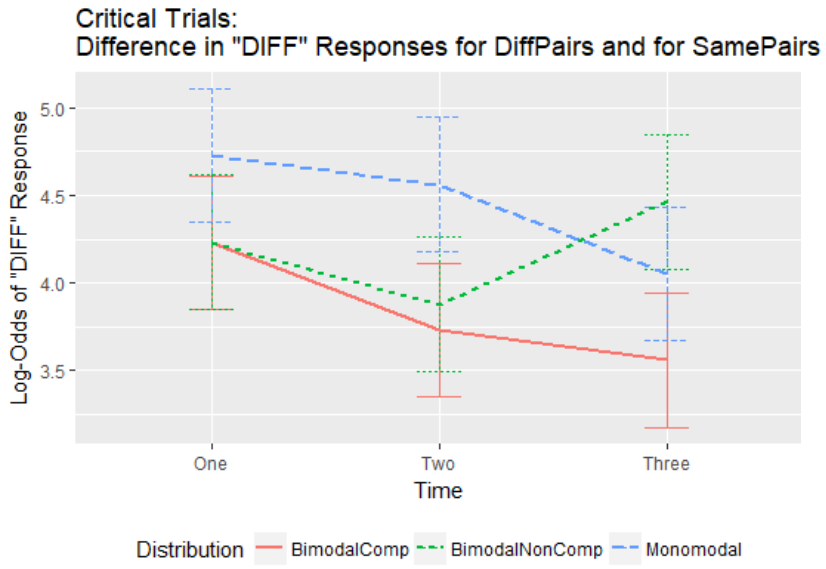

Figure 4. Sensitivity for critical trials; specifically, the difference (in log-odds) of participants responding that critical SamePairs are "different" compared to responding that critical DiffPairs are "different."

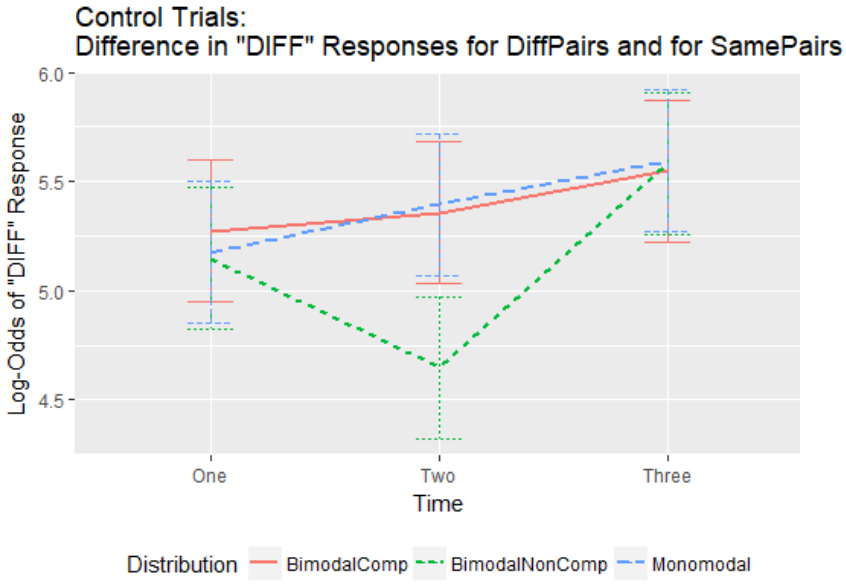

Figure 5. Sensitivity for control trials; specifically, the difference (in log-odds) of participants responding that control SamePairs are "different" compared to responding that control DiffPairs are "different."

6. Discussion. Interestingly, we find that the Bimodal-Comp group appears to have learned that critical items S_1a-S_4a and S_5a-S_8a belong to a single phoneme more quickly than the Monomodal group does. We can see this in the significantly lower sensitivity in the Bimodal-Comp group compared to the Monomodal group at ExposureTime Two. This section will briefly discuss why this may be the case.

In their one-step model of phoneme acquisition, Dillon et al. (2012) give the machine learner knowledge of the relevant conditioning environment which differentiates each pair of Inuktitut allophones. However, they acknowledge that this is not how a language learner acquires phonological rules, and that exactly *how* language learners determine which environments are relevant is still unanswered. We believe that the behavior of the Bimodal-Comp group with respect to the Monomodal group sheds light on this process. The Bimodal-Comp group coming to the conclusion that there is a single category more quickly than the Monomodal group can be explained if we model early acquisition as a process in which learners begin with the assumption that allophonic alternations exist, and search through their hypothesis space for possible conditioning environments. That is, they specifically search for *subsets* of sets with the initial hypothesis that there is more than one subset. Only when no conditioning environment which explains the learner's input is found does the learner settle on the hypothesis that there is no allophonic alternation. Therefore, the Bimodal-Comp group in this study is actually aided by the fact that there exists an easily-discoverable, (somewhat[6]) phonetically-natural allophonic alternation: [ɕ] occurs after [i] and [ʂ] occurs after [u]. After finding this conditioning environment, the Bimodal-Comp learner quickly settles on the hypothesis that [ɕ] and [ʂ] are allophones of a single phoneme. On the other hand, the Monomodal group may still be entertaining and testing various hypotheses regarding possible conditioning environments, and so does not settle on the hypothesis that there is just one post-alveolar fricative phoneme as quickly as the Bimodal-Comp group does.

7. Conclusion. This study finds support for a one-step model of phoneme acquisition, and successfully replicates results from Noguchi (2016). At no point during the learning trajectory mapped in this study did the Bimodal-Comp group exhibit higher sensitivities than the Monomodal group, and at ExposureTime Two (which corresponded to hearing 192 critical stimuli and about 10 minutes of training) exhibited significantly *lower* sensitivities than the Monomodal group. At Timepoint Three (which corresponded to hearing 288 critical stimuli and about 15 minutes of training), the Bimodal-Comp and Monomodal groups numerically appear to pattern together, with the Bimodal-Comp group having significantly lower sensitivites than the Bimodal-NonComp group. Additionally, this experiment makes the unexpected finding that the Bimodal-Comp group appears to learn more quickly that only one phoneme exists in the speech signal, in comparison to the Monomodal group. We suggest this may be because learners begin with the hypothesis that there exists some sort of allophonic alternation with some sort of conditioning environment, and only setttles on the alternate hypothesis, that there is none, after it has tested its entire search space for possible conditioning environments.

[6] See Noguchi (2016) for a discussion regarding the naturalness of this environment.

8. Appendix

Predictor	Coefficient	SE	Wald Z	*p*
(Intercept)	-0.095	0.328	-0.288	0.773
Distribution=*bimodalNonComp*	0.639	0.436	1.467	0.142
Distribution=*monomodal*	-0.186	0.442	-0.42	0.675
PairType=*same*	-4.230	0.309	-13.686	<0.001 ***
ExposureTime=*three*	0.006	0.452	0.013	0.990
ExposureTime =*two*	0.445	0.458	0.972	0.331
Interaction=*bimodalNonComp & same*	-0.003	0.383	-0.008	0.993
Interaction=*monomodal & same*	-0.497	0.443	-1.121	0.262
Interaction=*bimodalNonComp & three*	-0.270	0.629	-0.428	0.668
Interaction=*monomodalNonComp & three*	0.330	0.617	0.534	0.593
Interaction=*bimodalNonComp & two*	-0.672	0.652	-1.031	0.303
Interaction=*monomodalNonComp & two*	-0.431	0.634	-0.68	0.496
Interaction=*same & three*	0.672	0.387	1.736	0.083
Interaction=*same & two*	0.498	0.387	1.287	0.198
Interaction=*bimodalNonComp & same & three*	-0.903	0.547	-1.65	0.099
Interaction=*monomodalNonComp & same & three*	0.002	0.560	0.004	0.997
Interaction=*bimodalNonComp & same & two*	-0.143	0.533	-0.267	0.789
Interaction=*monomodalNonComp & same & two*	-0.334	0.599	-0.558	0.577

Table 4: Summary of results of GLMM for critical trials.

References

Boersma, Paul, Paola Escudero Escudero, and Rachel Hayes. 2003. Learning abstract phonological from auditory phonetic categories: An integrated model for the acquisition of language-specific sound categories. *15th International Congress of Phonetic Sciences.* 1013-1016.

Cristià, Alejandrina, Grant L McGuire, Amanda Seidl, and Alexander L Francis. 2011. Effects of the distribution of acoustic cues on infants' perception of sibilants. *Journal of Phonetics* 39(3). 388-402.

Dillon, Brian, Ewan Dunbar, and William Idsardi. 2012. A single-stage approach to learning phonological categories: Insights from Inuktitut. *Cognitive Science* 1. 34.

Eilers, Rebecca E, William Gavin, and Wesley R Wilson. 1979. Linguistic experience and phonemic perception in infancy: A crosslinguistic study. *Child Development* 14-18.

Eimas, Peter D, Siqueland, Peter Jusczyk, and James Vigorito. 1971. Speech perception in infants. *Science* 171(3968). 303-306.

Escudero, Paola, Titia Benders Benders, and Karin Wanrooij. 2011. Enhanced bimodal distributions facilitate the learning of second language vowels. *The Journal of the Acoustical Society of America* 130(4). EL206-EL212.

Feldman, Naomi, Thomas Griffiths Griffiths, and James Morgan. 2009. Learning phonetic categories by learning a lexicon. *Cognitive Science Society.*

Guenther, Frank H, and Marin N Gjaja. 1996. The perceptual magnet effect as an emergent property of neural map formation. *The Journal of the Acoustical Society of America* 100(2). 1111-1121.

Gulian, Margarita, Escudero Paola, and Paul Boersma. 2007. Supervision hampers distributional learning of vowel contrasts. *16th International Congress of Phonetic Sciences.* Saarbrücken, Germany: Saarland University. 1893-1896.
Hayes-Harb, Rachel. 2007. Lexical and statistical evidence in the acquisition of second language phonemes. *Second Language Research* 23 (1): 65-94.
Hothorn, Torsten, Frank Bretz, Peter Westfall, Richard M Heiberger, Andre Schuetzenmeister, and Susan Scheibe. 2017. *Package 'multcomp'.*
Kuhl, Patricia K. 2004. Early language acquisition: cracking the speech code. *Nature Reviews Neuroscience* 5 (11): 831-843.
Kuhl, Patricia K, Erica Stevens, Akiko Hayashi, Toshisada Deguchi, Shigeru Kiritani, and Paul Iverson. 2006. Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science* 9 (2).
Kuhl, Patricia K, Karen A Williams, Francisco Lacerda, Kenneth N Stevens, and Björn Lindblom. 1992. Linguistic experience alters phonetic perception in infants by 6 months of age. *Science* 255 (5044): 606-608.
Maye, Jessica, and LouAnn Gerken. 2000. Learning phonemes without minimal pairs. *24th Annual Boston University Conference on Language Development.*
Maye, Jessica, and LouAnn Gerken. 2001. Learning phonemes: How far can the input take us." *25th Annual Boston University Conference on Language Development.* Somerville, MA: Cascadilla Press.
Maye, Jessica, Janet F Werker, and LouAnn Gerken. 2002. Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition* 82 (3): B101-B111.
Moeng, Emily. Forthcoming. *The Acquisition of Phonetic Categories.* Chapel Hill, NC: The University of North Carolina at Chapel Hill.
Noguchi, Masaki. 2016. *Acquisition of allophony from speech input by adult learners.* Vancouver: The University of British Columbia.
Ong, Jia Hoong, Denis Burnham, and Paola Escudero. 2015. Distributional learning of lexical tones: a comparison of attended vs. unattended listening. *PLOS One* 10 (7): e0133446.
Peperkamp, Sharon, Michèle Pettinato Pettinato, and Emmanuel Dupoux. 2003. Allophonic variation and the acquisition of phoneme categories. *27th Annual Boston University Conference on Language Development.* Boston, MA: Cascadilla Press. 650-661.
Peperkamp, Sharon, Rozenn Le Calvez, Jean-Pierre Nadal, and Emmanuel Dupoux. 2006. The acquisition of allophonic rules: Statistical learning with linguistic constraints. *Cognition* 101 (3): B31-B41.
Werker, Janet F, and Richard C Tees. 1984. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development* 7 (1): 49-63.
Werker, Janet F, Ferran Pons, Christiane Dietrich, Sachiyo Kajikawa, Laurel Fais, and Shigeaki Amano. 2007. Infant-directed speech supports phonetic category learning in English and Japanese. *Cognition* 103 (1): 147-162.
Werker, Janet F, H. Henny Yeung, and Katherine A Yoshida. 2012. How do infants become experts at native-speech perception? *Current Directions in Psychological Science* 21 (4): 221-226.
Werker, Janet F, John HV Gilbert, Keith Humphrey, and Richard C Tees. 1981. Developmental aspects of cross-language speech perception. *Child Development* 349-355.